Budgets and the Extent of Noncompliance

The monitoring probability just defined, if faced by everyone of N identical sources, would provide an incentive for every source to choose to comply. But actual measurement would find the fraction $m\alpha$ in violation in every period. Expanding the observed fraction by the sample fraction implies, not surprisingly, that $\alpha \times 100$ percent of the sources could be expected to appear out of compliance if all sources were monitored in a given period.

The dollars spent on monitoring the fraction m of the sources would be NmM. Expressing this as a fraction of the cost of complete monitoring gives us: $m = \frac{NmM}{NM}$.

But suppose, as seems reasonable, that the available budget for monitoring would not support this much effort. How are the attainable fraction of sources in compliance and the budget required for that attainment related?

First, observe that the agency cannot proceed in a situation of limited budget ($r < m$) to announce a monitoring probability r applying to all sources, for then none would have an incentive to comply. Rather it must announce that the probability m applies to a suitably chosen subsect N_r, such that $\frac{N_r \cdot m \cdot M}{N \cdot M} = r$, the available budget.[4] That is $N_r = \frac{N \cdot r}{m}$. Then the expected fraction of sources out of compliance in a period can be inferred in two parts:

- those sources that do not expect to be monitored
- those sources that are monitored and are expected to be found out of compliance.

Denoting the expected fraction of sources in violation as ε, we obtain:

$$\varepsilon = \frac{N - N_r}{N} + \frac{N_r\alpha}{N}$$

$$= \frac{N - \frac{N.r}{M}}{N} + \frac{\frac{N.r.\alpha}{M}}{N}$$

$$= 1 - \frac{r}{M} + \frac{r\alpha}{M}$$

$$\text{or} \quad \varepsilon = 1 - (1-\alpha)\frac{r}{m}$$

Thus, if r is fixed, this ε is the smallest fraction of sources that can be expected to be in violation.

On the other hand, it may be of interest to determine what budget would be required to produce an expected violation fraction equal to some ε. This is given by: $r = \frac{(1-\varepsilon)}{(1-\alpha)}m$

for $1 \geq \varepsilon \geq \alpha > 0$.[6] Note for future reference that $\frac{\partial r}{\partial \varepsilon} = \frac{-m}{1-\alpha} < 0$.

That is, the required budget declines as the requirement on the fraction of sources in violation is relaxed.[5] Illustrative values of required budgets implied by assumed f, α and β, are provided in table 2.

Extending The Game To Repeated Periods Or Plays

In the agency-source game described above, the agency takes no account in its future plans of the discovered behavior of any source. Violators are fined, but that is the end of it. Another way of saying the same thing is that the only penalty for a discovered violation is a one-period fine, F. In the tax-compliance literature, models have been developed that make use of information on past behavior (as discovered by monitoring, of course) in defining either the future probability of monitoring or the future fine for a discovered violation or both. For example, Landsberger and Meijison 1982 allow both fine and probability to vary within a two state model, while Greenberg 1984 varies the monitoring probability only, but uses a three-state model.

Table 2: Budget Fractions Required To Obtain Chosen Compliance Levels

Allowed fraction not complying (ε)	Relative fine (f)	Type I error (α)	Required (Single play) budget (r_{sp})
0.2	1.2	0.05	.493
0.2	3.2	0.05	.263
0.1	1.2	0.05	.554
0.1	3.2	0.05	.296
0.2	1.2	0.01	.372
0.2	3.2	0.01	.196
0.1	1.2	0.01	.418
0.1	3.2	0.01	.220

These approaches to the construction of monitoring and enforcement systems have attracted attention because they appear to allow the attainment of given levels of compliance with the regulations in question using lower levels of spending by the authorities. Systems closely akin to repeated games with grouping of regulatees are found in practice as well. For example, the U.S. Internal Revenue Service is widely believed to use data from past audits to define the probability of a current year audit. And new regulations aimed at controlling ground water pollution (Fortuna and Lennett 1987) effectively group landfill operations for current monitoring purposes on the basis of past detection of problems.

A repeated monitoring game may conveniently be thought of as a markov process in which the states are groups among which the regulated parties move according to the results of particular plays of the embedded game. The probabilities of transition depend on the frequency with which the regulated parties are monitored when in each group. Discovery of a violation when in one group results in reassignment of the party to another group with different monitoring probability. In Greenberg's systems there are three groups. Being caught in violation while in group 1 results in reassignment to group 2 and a subsequent violation in group 2 would result in reassignment in perpetuity to group 3, in which auditing is constant (every year for an income tax system) and perpetual. A party monitored and found in compliance in group 2, however, is returned to group 1. Greenberg shows that in the absence of errors in the audit, and when the parties do not discount future costs, all parties in group 2 will comply and no party will ever be assigned to group 3. Monitoring probabilities in the first two groups can be small fractions of the levels necessary to stimulate compliance in a single play game situation. All parties in group 1 will find it optimal to violate because of their second chance in group 2.

There are several ways to interpret such systems. If one views the regulated parties as generally attempting to comply, grouping is simply a way of concentrating resources on those least capable of successfully doing so. An alternative, where compliance choice is taken to be an open question, is to see the variation of monitoring probability on the basis of past record as an integration of the two elements of the expected cost of noncompliance — size of fine and probability of its exaction. In this sense it corresponds to enforcement approaches involving higher fines for repeated offenses.

One may also interpret such extensions of single-play enforcement games as involving the added element of threat: If you don't do as I say you should, I will undertake a course of action neither of us really wants — checking up on you so frequently for such a long time that you will face a compliance "hell" and I will have to spend a considerable sum on your compliance alone. Behave in a reasonable fashion and I will offer occasional opportunities to relax. In the repeated game approach of Greenberg, for example, a "reasonable fashion" is implicitly defined as never failing two audits in a row. The model developed here extends Greenberg's three-state approach, and involves a different definition of the monitoring probabilities and the introduction of errors of inference.

Consider first, however, a simple version of the repeated play model without error. In what follows it is possible to relax the assumption that all sources are identical without adding to the expositional difficulty. When this is done, it will be convenient to define ρ as a monitoring probability lower than the smallest value of m_i, defined as above for each of the i sources ($m_i = \frac{1}{(1+f_i)(1-\beta) - \alpha f_i}$)

Assume for a moment that Greenberg's results hold, so that every source in group 2 has the incentive to comply. But none in group 1 do. Then group 3 remains empty if no source is assigned to it in the beginning, and the transition probability to group 3 from group 2 is zero. Define the monitoring probability in group 2 to be ρ for every source. Let the monitoring probability group 1 be expressed as a function of ε and ρ: $z(\varepsilon, \rho)$. In these circumstances the repeated play game produces movement of sources among groups 1 and 2 that is captured by a simple markov process with the following transition matrix (where G_i stands for group i):[6]

$$\begin{array}{ccc} & G_1 & G_2 \\ G_1 & 1-z(\varepsilon, \rho) & z(\varepsilon, \rho) \\ G_2 & \rho & 1-\rho \end{array}$$

It is easy to solve for the stationary probabilities Π_1, and Π_2, which can be intrepreted as the fractions of sources in the two groups in the long run:

$$\Pi_1 = \frac{\rho}{\rho + z(\varepsilon, \rho)}$$

$$\Pi_2 = \frac{z(\varepsilon, \rho)}{\rho + z(\varepsilon, \rho)}$$

The reason for expressing z as a function of ε and ρ is that Π_1, must equal ε, where ε is the required upper limit on the noncompliance fraction. Thus:

$$\varepsilon = \frac{\rho}{\rho + z(\varepsilon, \rho)}$$

$$\text{or } z(\varepsilon, \rho) = \rho\left(\frac{1}{\varepsilon} - 1\right)$$

To guarantee $\bar{z} < 1$, we must have $\rho < \frac{\varepsilon}{1-\varepsilon}$, but to guarantee $z < m$ so that in fact the incentive is not to comply in group 1, it is necessary that $\rho < m\left(\frac{\varepsilon}{1-\varepsilon}\right)$, an obviously stricter requirement. If $\varepsilon = \Pi_1$, $1-\varepsilon = \Pi_2$ the fraction of sources being audited in the steady state condition is:

$$\varepsilon\ (\rho\ (\frac{1}{\varepsilon} - 1)) + (1 - \varepsilon\ )\rho = 2\ \rho(1 - \varepsilon\ )$$

This is also the budget fraction as defined above for the single play game.

That is: $\dfrac{N\cdot 2\rho(1-\varepsilon)\cdot M}{N\cdot M} = 2\rho(1-\varepsilon) = r.$

What, then, is the relation between the budget fractions required in single play and multiple play games in the absence of error? Above, the errorless single-play budget fraction, r_{sp}, was found to be $(1-\varepsilon\ )m$ (footnote 5). Calling the multiple-play budget fraction r_{mp}, it is seen that:

$$\frac{r_{sp}}{r_{mp}} = \frac{(1-\varepsilon)m}{(1-\varepsilon)2\rho} = \frac{m}{2\rho}$$

Therefore, the budget required to produce a compliance fraction at least $(1-\varepsilon)$ will be smaller for the multiple play game as long as $m > 2\rho$ or $\rho < \frac{m}{2}$ (Thus, $\rho < \min(\frac{m}{2}, \frac{m\varepsilon}{1-\varepsilon})$ is a rule guaranteeing both sensible transition probabilities and budget savings from going to the multiple play game. If $\varepsilon = 1/3$, the two arguments for min ( ) are equal, and for $\varepsilon < 1/3$. the requirement that $\rho < m\varepsilon/(1-\varepsilon)$ governs. Therefore, in situations with small allowed fraction of sources in violation, the ratio $\frac{r_{sp}}{r_{mp}}$ will be greater than or equal to $1/2(\frac{1-\varepsilon}{\varepsilon})$, independently of the other parameters, such as fine size.

Introducing Errors Of Inference Into The Repeated Game

If the monitoring instruments available to the agency are not perfect but display random error (or if the parties cannot control their own actions perfectly but may randomly violate regulations in spite of intending not to) there will again be probabilities both of missing actual attempts to evade the regulations and of incorrectly identifying violations by sources in or attempting to be in compliance. The major implication of this change in assumptions will be that group 3 can no longer be assumed to be empty even if the incentive to comply exists in group 2. Indeed, group 3 will be an absorbing state if it is treated as a perpetual punishment for those caught in violation in group 2. Eventually, becuase of false positives, all parties would be in group 3, and the budget requirement would be 1 if all sources in group 3 are audited every period.

The major alteration to the repeated game structure required to accomodate these errors of inference is the introduction of some possibility of escape from group 3. But to link escape to successful passing of a monitoring visit, while perhaps initially tempting, is dangerous. Consider, for example, a design in which a party in group 3 is monitored each period. If success in a monitoring visit led to release back to group 2, the release probability would be $1 - \alpha$, where α is the probability of falsely identifying a violation when compliance is occuring. The average time spent in group 3 by a party would be $1/(1 - \alpha)$. It will be shown below that for reasonable values of the other parameters, the escape probability must be much less than $1 - \alpha$.

Rather, the probability of escape should be tailored to induce compliance in group 2, given the error structure <u>and</u> the size of the fine levied for discovered violations. Therefore, we write the transition matrix for a repeated game, with three groups and audit errors as:

$$
\begin{array}{cccc}
 & G_1 & G_2 & G_3 \\
G_1 & 1-z(.)(1-\beta) & z(.)(1-\beta) & 0 \\
G_2 & \rho(1-\alpha) & 1-\rho & \rho\alpha \\
G_3 & 0 & E & 1-E
\end{array}
$$

where, as before, ρ is fixed arbitrarily within an upper limit, and $z(.)$ will be chosen to maintain $\Pi_1 = \varepsilon$, the fraction of sources in violation, and will be a function of E, ρ, α, β, as well as ε. In what follows, ρ is taken to equal m/k, with k an arbitrary constant. α is the probability of a false positive and β the probability of a false negative, as above.

In this setting it is tedious but straightforward to derive the following results:

(i) $z(.) = \frac{E\rho(1-\alpha)(1-\varepsilon)}{\varepsilon(1-\beta)(E+\rho\alpha)}$ in order that $\Pi_1 = \varepsilon$

(ii) Then, $\bar{\Pi}_2 = \frac{E(1-\varepsilon)}{E+\rho\alpha}$ and

$$\Pi_3 = \frac{\alpha\rho(1-\varepsilon)}{E+\rho\alpha}$$

(iii) $r = \varepsilon z + \frac{E(1-\varepsilon)\rho}{E+\rho\alpha} + \frac{\alpha\rho(1-\varepsilon)A_3}{E+\rho\alpha}$

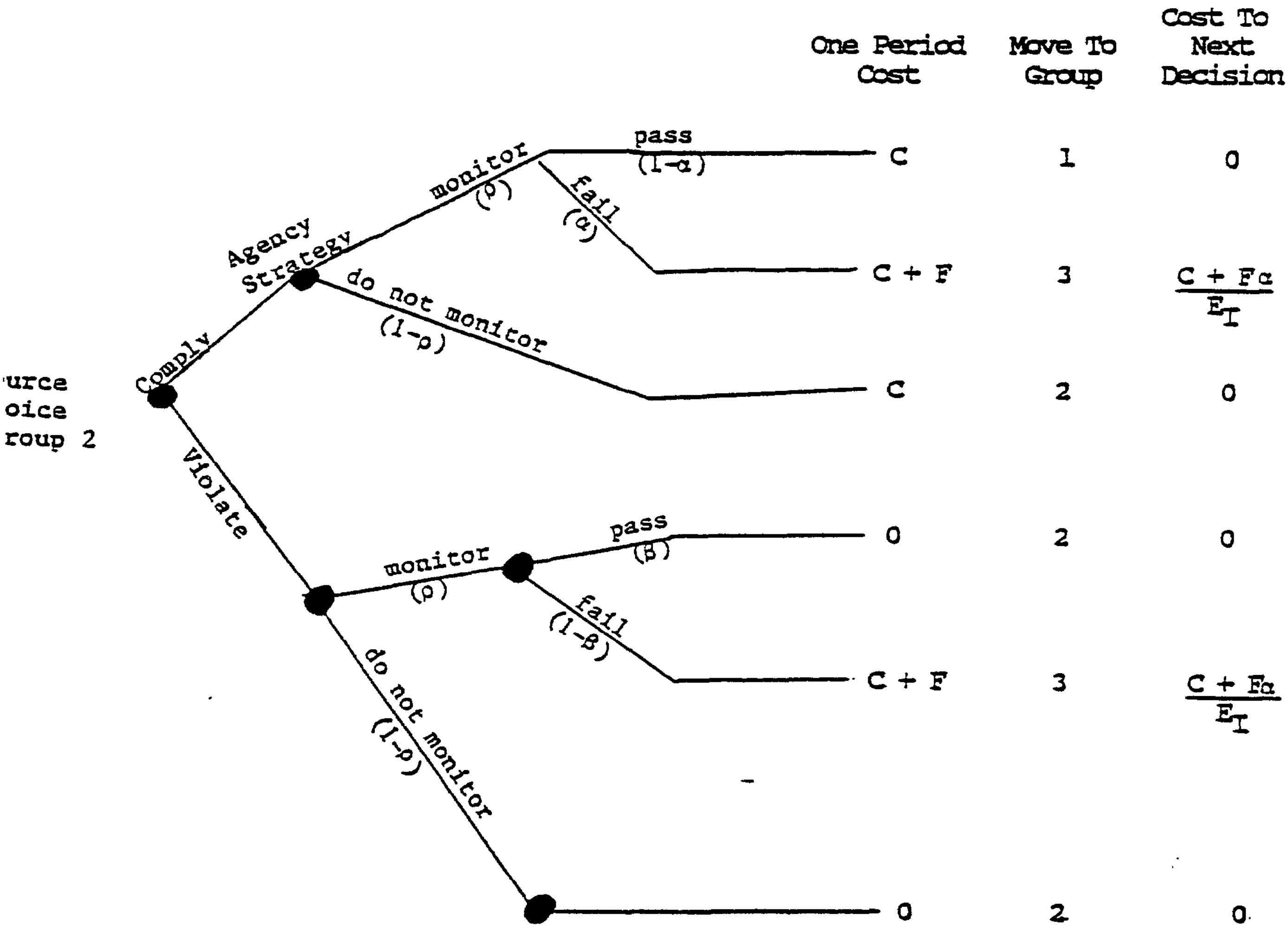
Where A_3 is the monitoring probability in group 3 and can be as low as is consistent with giving each source in that group the incentive to comply in every period. The results reported below generally involve assuming that $A_3 = 1$

(iv) Since $z = \frac{E\rho(1-\alpha)(1-\varepsilon)}{\varepsilon(1-\beta)(E+\rho\alpha)}$ and $\varepsilon z = \frac{E\rho(1-\alpha)(1-\varepsilon)}{(1-\beta)(E+\rho\alpha)}$

$\frac{\partial r}{\partial \varepsilon} < 0$, for r can be written as $(1-\varepsilon)[K]$, where $K > 0$ for

meaningful values of α, β, E, ρ, and A_3, and $\frac{\partial K}{\partial \varepsilon} = 0$

It is possible to derive an upper limit for E, call it E_T, that assures a source will have the incentive to comply in group 2. This derivation may be motivated by displaying the decision tree facing a source in group 2:



In this situation, the expected cost of compliance is:

$$E(\text{compliance}) = \rho\left[(1-\alpha)C + \alpha(C+F) + \alpha\left(\frac{C+F\alpha}{E_I}\right)\right] + (1-\rho)C;$$

While the expected cost of violation is: $E(\text{violation}) = \rho(1-\beta)\left[(C+F) + \frac{C+F\alpha}{E_I}\right]$;

Requiring that E(compliance) < E (violation) implies that:

$$E_I < \frac{\rho(1+f\alpha)(1-\beta-\alpha)}{1+\rho(f\alpha-(1-\beta)(1+f))}$$

Consistent values of ρ, f, α, β, and ε are used in table 3 to derive values of E_I. It may seem counter intuitive that as fines grow relative to avoidable costs of compliance, the upper limit on E_I falls (the time spent in group 3 grows). This occurs because as fines are increased, m and hence, ρ, the monitoring probability for group 2 both fall. This reduces the chance of

TABLE 3: CHARACTERISTICS OF THE REPEATED-PLAY MONITORING GAME, INCLUDING RELATIVE BUDGET SIZE, AS FUNCTIONS OF BASIC PARAMETERS

Required fraction in non-compliance ($\varepsilon = \Pi_1$)	Relative fine (f)	Type I error (α)	Type II error (β)	K ($\rho = \frac{m}{k}$)	Escape probability from group 3 (E_I)	Size of group 2 (Π_2)	Size of group 3 (Π_3)	Multiple play budget (r_{mp})	Single play budget (r_{sp})	Budget ratio (r_{sp}/r_{mp})
0.2	1.2	0.05	0.2	40	0.012	0.754	0.044	0.071	0.493	6.96
0.2	3.2	0.05	0.2	40	0.007	0.757	0.043	0.062	0.263	4.21
0.1	1.2	0.05	0.2	40	0.019	0.849	0.050	0.080	0.554	6.91
0.1	3.2	0.05	0.2	40	0.007	0.851	0.049	0.071	0.296	4.18
0.2	1.2	0.01	0.01	40	0.012	0.797	0.008	0.024	0.372	15.8
0.2	3.2	0.01	0.01	40	0.006	0.794	0.008	0.018	0.196	11.1
0.1	1.2	0.01	0.01	40	0.012	0.890	0.008	0.026	0.418	15.8
0.1	3.2	0.01	0.01	40	0.006	0.892	0.008	0.019	0.220	11.6
0.2	1.2	0.05	0.2	200	0.002	0.753	0.047	0.052	0.493	9.48
0.2	3.2	0.05	0.2	200	0.001	0.759	0.043	0.046	0.263	5.72

getting into group 3 at all, an effect that is balanced by the extension of expected time there. This demonstrates the central place in this system of the threat of residence in group 3 and the subsidiary role of the size of the fine levied for violation.

It is possible to substitute E_I into the expression for the budget, r, and thus to obtain an expression for the budget in terms of the underlying parameters, consistent with the requirements (i) that the proportion of violators be kept to ε; and (ii) that parties in group 2 have an incentive to comply. Unfortunately, the resulting expression is rather messy and does not bear any simple relation to the budget expression for the single play version of the game with error. Thus we obtain for r_{mp} (with error):

$$r_{mp} \text{ (with error)} = \frac{(1-\varepsilon)\rho\left[\left(\frac{\rho(1+f\alpha)(1-\beta-\alpha)}{1-\rho/m}\right)(2-\alpha-\beta)+\alpha(1-\beta)\right]}{\left(\frac{\rho(1+f\alpha)(1-\beta-\alpha)}{1-\rho/m}+\rho\alpha\right)(1-\beta)}$$

$$= \frac{(1-\varepsilon)\rho\left[E_I(2-\alpha-\beta)+\alpha(1-\beta)\right]}{(E_I+\rho\alpha)(1-\beta)}$$

where monitoring in group 3 is taken to be certain; $A_3 = 1$ in earlier notation.

To provide some feel for the budgetry advantage of using the threat implicit in the multiple-play and grouping approach to the enforcement game, however, table 3 also shows budget values for two different allowed proportions of violators, two levels of fines in relation to avoidable compliance costs, and two sets of error probabilities. The ratio of single-play to multiple-play budgets, parameter choices equal, is a measure of the improvement bought by introducing record-dependent monitoring probabilities. Note that with small errors, the ratio is over 15 to 1. Or, said another way, for this range of parameter choices, the budget required for a multiple-play scheme would be between 6 and 25 percent of a single-play approach designed to achieve the same compliance results.

Several other observations about table 3 are in order:

- As expected, <u>cet. par.</u> the required budget is lower when the allowed fraction of parties in violation is higher.
- But notice that the <u>ratio</u> of the single play to multiple-play budgets appears to be independent of the allowed fraction in violation. In fact, the small differences found in table 3 reflect rounding errors, for it is straightforward to show that in the ratio, all terms involving ε cancel.
- When the error structure is improved (both α and β reduced) budget requirements are reduced, other specifications remaining the same.
- A given increase in the size of the fine relative to avoidable cost has a larger effect on required budget in the single-play model than it does in the repeated game. This difference is more pronounced the less favorable the error structure.
- Budget size is affected by choice of the arbitrary constant, K, used to translate m (the compliance-inducing monitoring probability) into (the monitoring probability actually applied to group 2). It appears that simply by choosing K larger and larger (monitoring less and less in group 2) the budget can be driven closer and closer to zero. But this is an illusion created by the small range covered in the table. In fact, as K goes to infinity, the stationary probabilities Π_2 (fraction of sources in group 2) and Π_3 (fraction of sources in group 3) can be approximated by:

$$\Pi_2 \doteq \frac{1-\varepsilon}{1-\alpha}$$

$$\Pi_3 \doteq \frac{\alpha(1-\varepsilon)}{1+\alpha}$$